A New Class of Nonseparable Space-time Covariance Models

Thaís C.O. Fonseca and Mark F.J. Steel*

Abstract

The aim of this work is to construct nonseparable, stationary covariance functions for processes that vary continuously in space and time. Stochastic modeling of phenomena over space and time is important in many areas of applications such as environmental sciences, agriculture and meteorology. But choice of an appropriate model can be difficult as one must take care to use valid covariance structures. A common choice for the process is a product of purely spatial and temporal random processes. In this case, the resulting process possesses a separable covariance function. Although these models are guaranteed to be valid, they are severely limited, since they do not allow space-time interactions. In this work we propose a general and flexible way of constructing valid nonseparable covariance functions derived through mixing over separable covariance functions. The proposed model allows for different degrees of smoothness across space and time and long-range dependence in time. Moreover, our proposal has as particular cases several covariance models proposed in the literature such as the Matérn and the Cauchy Class. We use a Markov chain Monte Carlo sampler for Bayesian inference and apply our modeling approach to the Irish wind data.

Key words: Bayesian Inference; Irish wind data; Mixtures; Spatiotemporal modeling.

1 INTRODUCTION

Stochastic modeling of phenomena over space and time is in great demand in many areas of applications such as environmental sciences, agriculture and me-

*Thaís C.O. Fonseca is a Ph.D. student (Email: T.C.O.Fonseca@warwick.ac.uk) and Mark Steel is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: M.F.Steel@stats.warwick.ac.uk). Thaís Fonseca acknowledges financial support from the Center for Research in Statistical Methodology (CRiSM).

teorology. The aim of this work is to introduce a way of constructing nonseparable covariance functions for processes that vary continuously in space and time. Suppose that $(s,t) \in D \times T$, $D \subseteq \Re^d$, $T \subseteq \Re$ are space-time coordinates that vary continuously in $D \times T$ and we seek to define a spatiotemporal stochastic process $\{Z(s,t) : s \in D; t \in T\}$. In order to specify this process we need to determine the space-time covariance structure $C(s_1, s_2; t_1, t_2)$, for $s_1, s_2 \in D$ and $t_1, t_2 \in T$. In practice, it is often necessary to consider simplifying assumptions such as stationarity, isotropy, Gaussianity and separability. In what follows, we assume $\mathrm{Var}(Z(s,t)) < \infty$, for all $(s,t) \in D \times T$ and stationary covariance functions, that is, $\mathrm{Cov}(Z(s_0,t_0); Z(s_0+s, t_0+t)) = C(s,t),\ s \in D, t \in T$ depends on the space-time lag (s,t) only, for any $s_0 \in D,\ t_0 \in T$. Choice of an appropriate model can be difficult as one must take care to use valid covariance structures, that is, for any $(s_1,t_1), ..., (s_m,t_m)$, any real $a_1, ..., a_m$, and any positive integer m, C must satisfy $\sum_{i=1}^m \sum_{j=1}^m a_i a_j C(s_i - s_j, t_i - t_j) \geq 0$, as this is the covariance function $\mathrm{Var}(\sum_{i=1}^m a_i Z(s_i, t_i))$ of real linear combinations of $Z(s,t)$. A common choice for the process $\{Z(s,t), (s,t) \in D \times T\}$ is $Z(s,t) = Z_1(s) Z_2(t),\ (s;t) \in D \times T$, where $\{Z_1(s) : s \in D\}$ is a purely spatial random process with covariance function $C_1(s)$ and $\{Z_2(t) : t \in T\}$ is a purely temporal random process with covariance function $C_2(t)$. The processes $Z_1(s)$ and $Z_2(t)$ are uncorrelated. The resulting process $Z(s,t)$ possesses a separable covariance function given by

$$C(s,t) = C_1(s) C_2(t),\ (s,t) \in D \times T, \tag{1}$$

where C_1 and C_2 are valid covariance functions. The validity of the resulting covariance function in (1) comes from the property that sums, products, convex combinations and limits of positive definite functions are positive definite. Separability is a convenient property since the covariance matrix can be expressed as a Kronecker product of smaller matrices that come from the purely temporal and purely spatial processes. Thus, determinants and inverses are easily obtained providing a potentially large computational benefit. Although the assumption of separable processes in time and space is very convenient, it is usually unrealistic. Cressie and Huang (1999) discusses some shortcomings of separable models. While these models are guaranteed to be valid (satisfy the positive definiteness condition), they are severely limited, since they do not allow space-time interactions. Stein (2005) points out that separable models have lack of smoothness away from the origin, that is, small changes in the location of observations can lead to large changes in the correlation between certain linear combinations of observations. This will not happen for analytic functions such as $c \exp(at^2)$ that are infinitely differentiable, but these functions are not adequate for physical processes and the Matérn Class is suggested as an alternative for differentiable processes. Some recent advances were made in developing valid nonseparable models. Carrol et al. (1997), Cressie and Huang (1999), Gneiting (2002) and Stein (2005) all suggest ways of constructing nonseparable covariance models. Carrol et al. (1997) propose a nonseparable spatiotemporal model to reconstruct ozone surfaces and estimate the population exposure in Harris county. They do not show analytically that the covariance function is positive definite and concerns about the validity of the model were raised in comments by Cressie (1997) and Guttorp et al. (1997). Cressie and Huang (1999) introduce new classes of nonseparable, stationary covariance functions that allow for space-time interaction but the approach is restricted to a small class of functions for which a closed-form solution to a d-variate Fourier integral is known. Gneiting (2002) proposes a new class of valid covariance models. The same approach as Cressie and Huang (1999) is adopted but the Fourier integral limitation is avoided. A criterion for positive definiteness is proposed to validate covariance functions and it is used to show that some of the space-time covariance functions presented by Carrol et al. (1997) and Cressie and Huang (1999) are not valid. Stein (2005) considers stationary covariance functions and discusses what space-time covariances imply about the corresponding processes. The author points out that the examples provided by Cressie and Huang (1999) are analytic, that is, do not have lack of smoothness away from the origin, but the general approach can yield covariance functions without this property. Another comment is that the nonseparable functions proposed by Gneiting (2002) are possibly not smoother along their axes than at the origin. An example of space-time covariance functions that can achieve any degree of differentiability in space and in time is provided but the general approach does not provide explicit expressions for the covariance functions. In this work we propose a general and flexible way of constructing valid nonseparable covariance functions derived through mixing over separable covariance functions. Section 2 provides the general mixing approach that guarantees positive definiteness for the class. Section 3 provides the proposed general class of covariance models that allows for different degrees of smoothness across space and time. The purely temporal process can achieve different degrees of smoothness while the purely spatial process can possess a covariance function with the same differentiability properties as the Matérn Class. Moreover, for any given location $s_0 \in D$, the purely temporal process $Z(s_0, .)$ can have long-range dependence in time. Inference on these models will be conducted from a Bayesian perspective through Markov chain Monte Carlo (MCMC) methods, as described in Section 4. Code for the implementation of this inference is freely available on `http://www.warwick.ac.uk/go/msteel/steel_homepage/software/`. An application to a well-known data set, the Irish wind data, is provided in Section

5. The final section concludes. Proofs will be deferred to Appendix B without mention in the text.

2 MIXTURE REPRESENTATION

Let (U, V) be a bivariate nonnegative random vector with distribution $\mu(u, v)$ and independent of $\{Z_1(s), s \in D\}$ and $\{Z_2(t), t \in T\}$. Define the process

$$Z(s,t) = Z_1(s;U)Z_2(t;V),\ (s,t) \in D \times T \tag{2}$$

where $Z_1(s;u)$ is a purely spatial random process for every $u \in \Re_+$ with covariance function $C_1(s;u)$ that is a stationary covariance for $s \in D$ and every $u \in \Re_+$ and a measurable function of $u \in \Re_+$ for every $s \in D$. And $Z_2(t;v)$ is a purely temporal random process for every $v \in \Re_+$ with covariance function $C_2(t;v)$ that is a stationary covariance for $t \in T$ and every $v \in \Re_+$ and a measurable function of $v \in \Re_+$ for every $t \in T$. Then the corresponding covariance function of $Z(s,t)$ is a convex combination of separable covariance functions. It is valid (Ma, 2002, 2003, see) and generally non-separable, and is given by

$$C(s,t) = \int C_1(s;u)C_2(t;v)d\mu(u,v) \tag{3}$$

It reduces to a separable covariance function if U and V are independent. This covariance function is a mixture of separable covariance functions and conditional on $U = u_0$ and $V = v_0$ the process $Z(s,t)$ possesses a separable covariance $C_1(s;u_0)C_2(t;v_0)$. One benefit of the mixing approach is that it generates a large variety of valid nonseparable, spatiotemporal covariance models, by using appropriate choices of the mixing function and the purely spatial and temporal covariances. Moreover, it takes advantage of the well-known theory developed for purely spatial and purely temporal processes in the joint modeling of space-time interactions. Various authors derived covariance models using the mixing representation, leading to a number of overlapping models. The simplest special case of (3) is given by a discrete mixture

$$C(s,t) = p_{11}C_1(s)C_2(t) + p_{10}C_1(s)C_2(0) + p_{01}C_1(0)C_2(t) + p_{00}C_1(0)C_2(0),$$

where $C_1(s;u) = C_1(su)$, $C_2(t;v) = C_2(tv)$ and $P(U = i, V = j) = p_{ij},\ i,j \in \{0,1\}$. A closely related model is the product-sum model of De Cesare et al. (2001) where some constraints are imposed in order to guarantee positive definiteness. Ma (2002, 2003) develops nonseparable covariances generated by using two approaches. One is a positive power mixture where $C_1(s;u) = C_1(s)^u$, $C_2(t;v) = C_2(t)^v$ and (U,V) is a nonnegative bivariate discrete random vector with probability function $\{p_{ij}, (i,j) \in Z_+^2\}$. The other is a scale mixture that considers a nonstationary version of (3), that is, $C_1(s_1, s_2; u) = C_1(s_1 u, s_2 u)$ depends on s_1 and s_2 and $C_2(t_1, t_2; v) = C_2(t_1 v, t_2 v)$ depends on t_1 and t_2.

Proposition 2.1 *Consider a bivariate nonnegative random vector (U, V) with joint moment generation function $M(.,.)$. If $\gamma_1 = \gamma_1(s)$ is a purely spatial variogram on D, $\gamma_2 = \gamma_2(t)$ is a purely temporal variogram on T and $C_1(s; u) = \exp(-\gamma_1 u)$ and $C_2(t; v) = \exp(-\gamma_2 v)$, then (3) becomes*

$$C(s,t) = M(-\gamma_1, -\gamma_2), \; (s,t) \in D \times T, \tag{4}$$

which is a valid spatiotemporal covariance function on $D \times T$.

The assumptions $C_1(s; u) = \exp(-\gamma_1 u)$ and $C_2(t; v) = \exp(-\gamma_2 v)$ are equivalent to $C_1(s; u)$ and $C_2(t; v)$ being valid covariance functions for any $u,\ v > 0$ (see e.g. Chilès and Delfiner, 1999, p. 66-67) and thus guarantee positive definiteness of the space-time covariance function. Another important feature in the model specification is the desirable flexibility of the class of models. In this work, we follow the specification (4) and propose a general way to define the nonnegative bivariate random vector (U, V) that leads to flexible nonseparable covariance functions with very useful properties.

3 A NONSEPARABLE COVARIANCE STRUCTURE

Proposition 3.1 *Consider X_0, X_1 and X_2, which are independent nonnegative random variables with finite moment generating functions M_0, M_1 and M_2, respectively and define $U = X_0 + X_1$ and $V = X_0 + X_2$. Let $C_1(s; u) = \sigma \exp\{-\gamma_1 u\}$ and $C_2(t; v) = \sigma \exp\{-\gamma_2 v\}$, with γ_1 and γ_2 as in proposition 2.1. Then the resulting space-time covariance function from (3) is*

$$C(s,t) = \sigma^2 M_0(-\gamma_1 - \gamma_2) M_1(-\gamma_1) M_2(-\gamma_2), \; (s,t) \in D \times T, \tag{5}$$

where σ^2 is the space-time variance.

Notice that if U and V are uncorrelated then the separable case is obtained and $C(s,t) = \sigma^2 M_1(-\gamma_1) M_2(-\gamma_2)$. The class generated as in Proposition 3.1 is a very flexible class since it allows for different parametric representations of space-time iterations, depending on the distributions of X_0, X_1 and X_2. As a consequence of the construction, any nonzero correlation between U and V will always be positive. We now consider building a general class of bivariate distributions for (U, V) using generalized inverse Gaussian (GIG) distributions, which are described in Appendix A. A gamma distribution with shape parameter λ and scale a (with mean λ/a) will be denoted as $\mathrm{Ga}(\lambda, a)$.

Theorem 3.1 *Consider $X_i \sim \mathrm{Ga}(\lambda_i, a_i)$ for $i = 0, 2$ and $X_1 \sim \mathrm{GIG}(\lambda_1, \delta, a_1)$, then the corresponding space-time covariance function generated through Proposition 3.1 is*

$$C(s,t) = \sigma^2 \left\{1 + \frac{\gamma_1 + \gamma_2}{a_0}\right\}^{-\lambda_0} \left\{1 + \frac{\gamma_1}{a_1}\right\}^{-\frac{\lambda_1}{2}} \frac{K_{\lambda_1}(2\sqrt{(a_1+\gamma_1)\delta})}{K_{\lambda_1}(2\sqrt{a_1\delta})} \left\{1 + \frac{\gamma_2}{a_2}\right\}^{-\lambda_2}, \tag{6}$$

where $K_\lambda(.)$ is the modified Bessel function of the second kind of order λ. Permitted parameter values are $\sigma^2 > 0$, $\lambda_0 \geq 0$, $a_0 > 0$, $\lambda_2 > 0$, $a_2 > 0$, and we allow for $a_1 > 0$, $\delta \geq 0$ if $\lambda_1 > 0$, whereas $a_1 > 0$, $\delta > 0$ if $\lambda_1 = 0$ and $a_1 \geq 0$, $\delta > 0$ if $\lambda_1 < 0$.

Notice that in Theorem 3.1 the structure derived in space is more complex than the one derived in time. If the main interest was the time dimension, we could have put a GIG distribution on X_2 instead to generate more complex structures in time (in principle, we could even use GIG distributions for both X_1 and X_2). When $a_1 = 0$ we use the asymptotic formula $K_{\lambda_1}(x) = 2^{\lambda_1 - 1}\Gamma(\lambda_1)x^{-\lambda_1}$ resulting in a inverse gamma moment generating function as will be illustrated in Model 2 (Subsection 3.1). In the representation (6), the parameter σ^2 is the space-time variance, that is, $\sigma^2 = C(0,0)$. The parameters a_1 and δ explain the rate of decay for the spatial correlation and a_2 has the same role in the temporal dimension. To avoid lack of identifiability in the model we fix $a_0 = 1$. Note that if $a_0 \neq 1$ the resulting class would not change but we would have a superfluous scale parameter. Contour plots of some spatiotemporal covariance function for this class is given in Figure 1.

It is important to measure separability in space time in the proposed model. We suggest to use the correlation between the variables U and V as an indication of interaction between space and time components. This correlation is given by

$$c = \frac{\lambda_0}{\sqrt{(\lambda_0 + V_1)(\lambda_0 + \lambda_2/a_2^2)}}, \tag{7}$$

where $V_1 = \mathrm{Var}(X_1)$ is defined in (18) for $a = a_1 > 0$, $\delta > 0$ and $\lambda = \lambda_1$. Thus, $0 \leq c \leq 1$ could be used as a measure of space-time interaction, with $c = 0$ indicating separability and $c = 1$ meaning high dependence between space and time. The parameter λ_0 plays the role of separability parameter and the separable case is obtained for $\lambda_0 = 0$. In this case we say that $X_0 = 0$, implying $U = X_1$ and $V = X_2$. On the other hand, if $\lambda_0 \to \infty$, $U \to V$ resulting in an extreme non-separable case. Plots of a separable ($c = 0$) and a nonseparable ($c = 0.98$) spatiotemporal covariance function for this class are given in Figure 2. Note from the plots of $\rho(s,t)/\rho(0,t) = C(s,t)/C(0,t)$ that the decay of the correlations in space is less rapid for larger differences in time, t. We can show

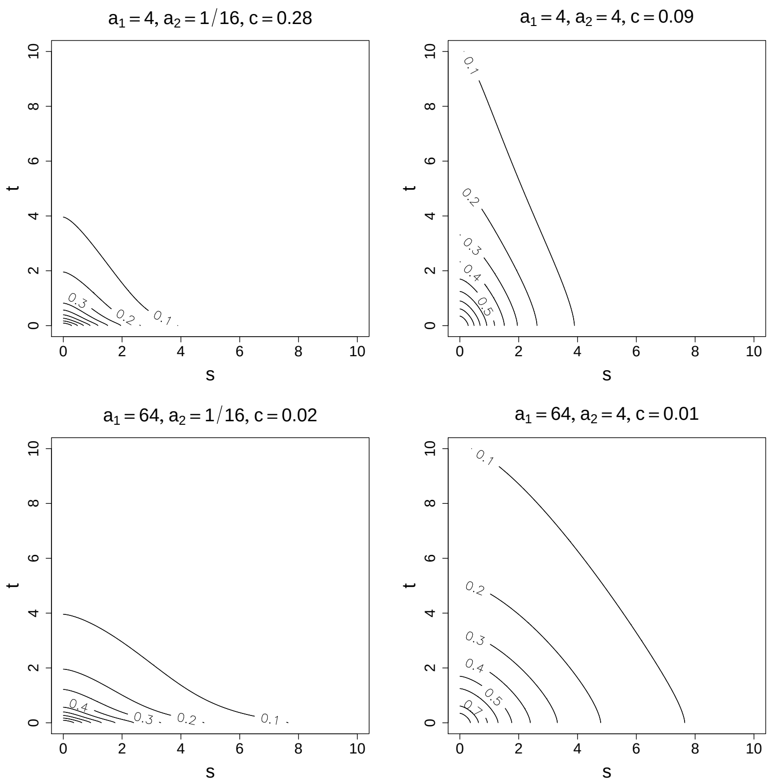


Figure 1: Contour plot of $C(s,t)$ in (6) for $\gamma_1(t) = ||s||^\alpha$, $\gamma_2(t) = |t|^\beta$, $\sigma^2 = 1$, $\delta = 1/16$, $\alpha = 1.5$, $\beta = 1.5$, $\lambda_0 = 1/2$, $\lambda_1 = 1/4$ and $\lambda_2 = 1/4$. The horizontal axis represents the spatial lag and the vertical axis, the temporal lag.

that whenever $\lambda_0 > 0$ the ratio $C(s,t)/C(0,t)$ is always a strictly increasing function of t for any s, so this behavior is a feature of the construction.

In the case where $\lambda_0 = 0$ and $\gamma_1 = ||s||^2$ the spatial margin is a generalization of the Matérn Class, proposed by Shkarofsky (1968) that allows two complementary positive scale parameters a_1 and δ. If also $a_1 = 0$ we obtain the Matérn Class as a particular case. For the temporal margin, if $\gamma_2 = |t|^\beta$ and the process is separable the resulting covariance function is in the Cauchy Class (Gneiting and Schlather, 2004). This class is the temporal margin obtained for most of the non-separable models proposed in the literature providing flexible power-law correlations that generalize stochastic models used in several fields. When $\lambda_0 \neq 0$ we have a model that is similar to the generalized Matérn ($a_1 \neq 0$) and to the Matérn ($a_1 = 0$) in the space dimension and similar to the Cauchy class in the time dimension but it also allows for space-time interactions.

Another important feature of the class (6) is that for any given $s_0 \in D$, the purely temporal process $Z(s_0, .)$ can have long-range dependence, a global characteristic associated with power law correlations. Consider $\gamma_2 = |t|^\beta$, if

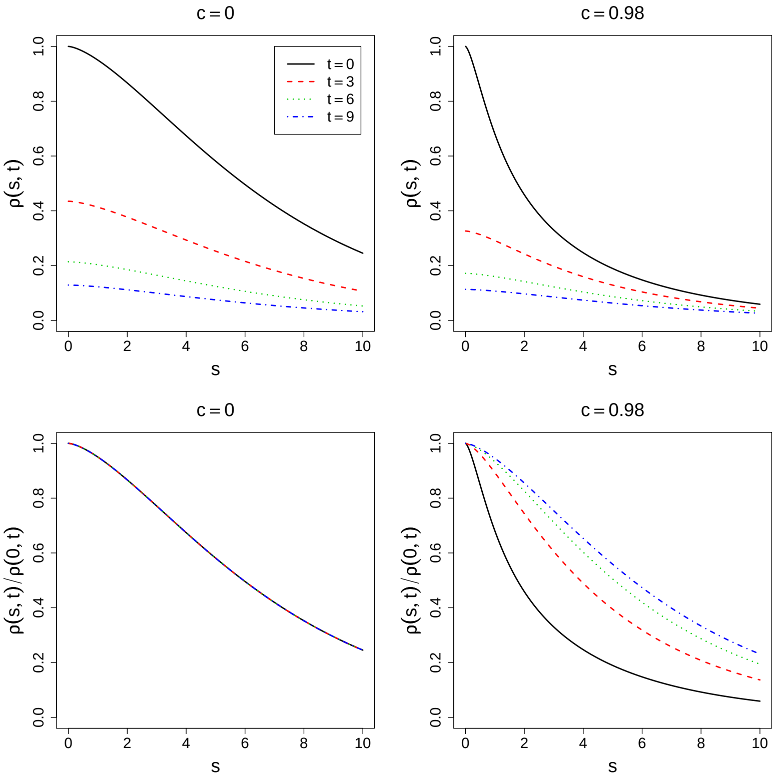


Figure 2: Plot of $\rho(s,t) = C(s,t)/C(0,0)$ and its normalized version for $\gamma_1(t) = ||s||^\alpha$, $\gamma_2(t) = |t|^\beta$, $\alpha = 1.5$, $\beta = 1.5$. The scales are $a_1 = 64$, $a_2 = 4$ and $\delta = 1/16$. For $c = 0$, we have taken $\lambda_0 = 0$, $\lambda_2 = \lambda_1 = 1$ and for $c = 0.98$, $\lambda_0 = 1/2$, $\lambda_2 = \lambda_1 = 1/4$. The horizontal axis represents the spatial lag.

$0 < \lambda_0 + \lambda_2 < 1/\beta$ then $C(0,t) \sim |t|^{-\beta(\lambda_0+\lambda_2)}$ as $t \to \infty$ and the purely temporal process is said to have long memory dependence. This characteristic implies that correlations between distant times decay much slower than for standard ARMA or Markov-type models. The closer $\lambda_0 + \lambda_2$ is to zero, the stronger the dependence of the process. The class also allows for short memory (random walk) when $\lambda_0 + \lambda_2 = 1/\beta$ and intermediate memory (antipersistent process) for $1/\beta < \lambda_0 + \lambda_2 < 2/\beta$. As a first step to provide smoothness properties of the process $\{Z(s,t) : s \in D,\ t \in T\}$, we study the behavior of $C(s,t)$ across space for a fixed time $t_0 \in T$ and across time for a fixed location $s_0 \in D$. Generally, $f^{(q)}$ will denote the q^{th} derivative of a function f. Consider the Taylor expansion of $\gamma_1(s)$ about 0 given by $\sum_{k=0}^{\infty} c_k ||s||^k$. Define l, the smallest power of $||s||$ in the Taylor expansion such that $c_l \neq 0$.

Proposition 3.2 *Under the conditions of Theorem 3.1,*
(a) the purely temporal process $\{Z(s_0,t) : t \in T\}$ at a fixed location $s_0 \in D$ is m times mean square differentiable if and only if $\gamma_2^{(2m)}(0)$ exists and is finite.

(b) When $a_1 \neq 0$, the purely spatial process $\{Z(s, t_0) : s \in D\}$ at a fixed time $t_0 \in T$ is m times mean square differentiable if and only if $\gamma_1^{(2m)}(0)$ exists and is finite. When $a_1 = 0$, the purely spatial process is m times mean square differentiable if and only if $2m < -l\lambda_1$ and $\gamma_1^{(2m)}(0)$ exists and is finite.

The proposed covariance model (6) allows for different degrees of smoothness across space and time obtained by choosing the parameter λ_1 and the functions γ_1 and γ_2. When $a_1 = 0$ (Matérn Class), the parameter $\lambda_1 < 0$ has a direct effect on the smoothness of the spatial process, and larger values of $-\lambda_1$ corresponds to smoother processes. The Matérn Class is very flexible, in the sense that the model allows for the degree of smoothness to be estimated from the data rather than restricted a priori. But some characteristics, such as no cusp at the origin and negative second derivative, required in *e.g.* turbulence applications, are not fulfilled by this class. On the other hand, the covariance function for the purely spatial process obtained when $a_1 \neq 0$ has no cusp at the origin, that is, $\frac{d}{ds}C(s, 0)$ goes to zero as s approaches zero. Furthermore, its second derivative always exists, is finite and is negative if $\gamma_1^{(2)}(0)$ exists and is finite. If we take $\gamma_1 = ||s||^\alpha$, $\gamma_2 = |t|^\beta$, $\alpha \in (0, 2]$ and $\beta \in (0, 2]$, we obtain the following.

Corollary 3.1 *Consider $\gamma_1 = ||s||^\alpha$, $\gamma_2 = |t|^\beta$, $\alpha \in (0, 2]$ and $\beta \in (0, 2]$ and conditions of Proposition 3.2. The temporal process is mean square continuous for $\beta \in (0, 2)$ and it is infinitely mean square differentiable for $\beta = 2$. The spatial process is mean square continuous for $\alpha \in (0, 2)$. When $\alpha = 2$, the process is m times mean square differentiable for $a_1 = 0$ and $-\lambda_1 > m$ and it is infinitely mean square differentiable for $a_1 \neq 0$.*

3.1 Parameterisation

We have specified a rich class of covariance structures in (6), and we now discuss useful parameterisations and interesting subclasses. As mentioned above, σ^2 is the space-time variance, and the parameters a_1 and δ are spatial scales, while a_2 is a scale parameter in the temporal dimension. We now introduce extra scale parameters in the variograms γ_1 and γ_2 by taking $\gamma_1(s) = ||s/a||^\alpha$ and $\gamma_2(t) = |t/b|^\beta$, where $\alpha \in (0, 2]$ and $\beta \in (0, 2]$. Note that these extra scale parameters do not change any of the results on smoothness or temporal dependence explained above.

We have to make restrictions with respect to the scale parameters, since the general class would now have $(a_0, a_1, a_2, \delta, a, b)$ as scales. As stated before, we fix $a_0 = 1$ because we already have scales in space and time and this extra one would be superfluous. Another reason for that choice is that a_0 would also influence the degree of separability in space and time. In particular, for $a_0 \to 0$

we would obtain $c \to 1$ (strong nonseparability) and when $a_0 \to \infty$ then $c \to 0$ (separability). By fixing a_0 we keep just λ_0 as a separability parameter. The same motivation leads us to consider (a, b) instead of (a_1, a_2, δ) as scales. Since (a_1, a_2, δ) also appear in the expression (7) for the degree of dependence in space and time, they would again be confounded with λ_0. On the other hand, (a, b) do not enter in this measure of separability. To avoid redundancy, Cressie and Huang (1999) decide to use the same scales as suggested here. The main difference is that in their example 5, for instance, the extra scales are actually necessary for the existence of a flexible separability parameter in the model while here (a_1, δ, a_2) and (a, b) are always redundant if used at the same time.

As discussed before, the functions $\gamma_1(s)$ and $\gamma_2(t)$ determine the smoothness of the random process $Z(s, t)$. It is difficult to decide whether to fix β at a particular value or not since for $\beta = 2$ the process is infinitely smooth and for $\beta < 2$ it is not even once mean square differentiable. The same holds for α when $a_1 \neq 0$. Depending on the application it might be appropriate to estimate these parameters. If we do set $\alpha = 2$ then this generates the subclass that gives the Matérn covariance function in space when $a_1 = 0$ and under separability. The Matérn class is important for spatial applications since, besides the scale parameter, it also has a smoothness parameter controlling the differentiability of the random field (see Corollary 3.1).

In the following we present some interesting parametric families of spatiotemporal covariance functions adhering to these parameter restrictions. Throughout, $a_2 = 1$ and we take $a > 0$ as the scale in space and $b > 0$ as the scale in time. The models below differ in how we constrain δ and a_1.

Model 1

An interesting general model is obtained by setting $\delta = a_1$ which means that δ is now a concentration parameter. The resulting covariance function is

$$C(s,t) = \sigma^2 \left\{1 + ||s/a||^\alpha + |t/b|^\beta\right\}^{-\lambda_0} \left\{1 + \frac{||s/a||^\alpha}{\delta}\right\}^{-\lambda_1/2} \frac{K_{\lambda_1}\left(2\delta\sqrt{1+\frac{||s/a||^\alpha}{\delta}}\right)}{K_{\lambda_1}(2\delta)} \left\{1 + |t/b|^\beta\right\}^{-\lambda_2}. \tag{8}$$

The dependence in space and time is given by (7) where $V_1(\lambda_1, \delta) = \mathrm{Var}(X_1)$ and $X_1 \sim \mathrm{GIG}(\lambda_1, \delta, \delta)$ and $a_2 = 1$. If $\lambda_0 = 0$, we have independence between U and V and the purely spatial covariance is a generalized Matérn for $\alpha = 2$ while the purely temporal covariance is in the Cauchy Class.

Model 1a

As a special case of Model 1, consider $\lambda_1 = -1/2$ so that $X_1 \sim$

InvGaussian(δ, δ) and the space-time covariance function is

$$C(s,t) = \sigma^2(1 + ||s/a||^\alpha + |t/b|^\beta)^{-\lambda_0}(1 + |t/b|^\beta)^{-\lambda_2} \exp\left\{-2\delta\left[\sqrt{1 + \frac{||s/a||^\alpha}{\delta}} - 1\right]\right\}. \tag{9}$$

If $\lambda_0 = 0$ the purely spatial covariance is a shifted version of the exponential covariance function for $\alpha = 2$ and the purely temporal covariance is in the Cauchy Class.

Model 2

Consider $\lambda_1 < 0$, $a_1 = 0$ and $\delta = 1$. Then, X_1 has an inverse gamma distribution $X_1 \sim \text{InvGa}(\nu = -\lambda_1, 1)$ and the space-time covariance function is

$$C(s,t) = \sigma^2 \left\{1 + ||s/a||^\alpha + |t/b|^\beta\right\}^{-\lambda_0} \frac{(2||s/a||)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu\left(2||s/a||^{\alpha/2}\right)\left\{1 + |t/b|^\beta\right\}^{-\lambda_2}. \tag{10}$$

As the variance of X_1 does not exist (unless $\nu > 2$ is imposed through the prior), the dependence in space and time is now measured by

$$\tilde{c} = \frac{\lambda_0}{\sqrt{(\lambda_0 + \tilde{V}_1(\nu))(\lambda_0 + \lambda_2)}}, \tag{11}$$

where $\tilde{V}_1(\nu) = (Q(0.75;\nu) - Q(0.25;\nu))^2$ and $Q(x;\nu)$ is the quantile corresponding to $100x\%$ of X_1. Under independence of U and V the purely spatial covariance is in the Matérn Class if we take $\alpha = 2$ and the purely temporal covariance is in the Cauchy Class. Finally, if we use $\lambda_0 = 0$ in combination with $\nu = 1/2$ we generate a powered exponential covariance structure in space.

Model 3

Consider $\lambda_1 > 0$ and the restrictions $\delta = 0$ and $a_1 = 1$. Then, $X_1 \sim \text{Ga}(\lambda_1, 1)$ and the resulting space-time covariance function is

$$C(s,t) = \sigma^2 \left\{1 + ||s/a||^\alpha + |t/b|^\beta\right\}^{-\lambda_0} \left\{1 + ||s/a||^\alpha\right\}^{-\lambda_1} \left\{1 + |t/b|^\beta\right\}^{-\lambda_2}. \tag{12}$$

In this example, the random vector (U, V) has the bivariate gamma distribution of Cheriyan-Ramabhardran and a closely related model is Example 6 of Ma (2002) where a nonstationary function is considered. If $\lambda_0 = 0$, U and V are independent gamma random variables and both purely spatial and purely temporal covariances are in the Cauchy Class. Nonseparability can be measured as in (7) with $V_1 = \lambda_1$ and $a_2 = 1$.

3.2 Including a Nugget Effect

In practice, it is often useful to consider discontinuities at the origin (nugget effect), to capture measurement error and small scale variation. This can be done in a natural way, using the mixture construction. Let us focus on a spatial nugget effect in what follows, but a temporal nugget effect can be dealt with in a similar fashion. Instead of (2) we consider $Z^*(s,t) = Z_1^*(s;U)Z_2(t;V)$, $(s,t) \in D \times T$, where $Z_1^*(s;U) = \sqrt{1-\tau^2}Z_1(s;U) + \tau\epsilon(s)$ with $\{\epsilon(s) : s \in D\}$ a process with zero mean, variance one and $\mathrm{Cov}(\epsilon(s_1), \epsilon(s_2)) = 0$ if $s_1 \neq s_2$, and $0 < \tau < 1$. We assume that this process is uncorrelated with the purely spatial and purely temporal processes. Under the conditions of Proposition 3.1, the resulting covariance function is then

$$C^*(s,t) = \sigma^2 M_0(-\gamma_1 - \gamma_2)M_1^*(-\gamma_1)M_2(-\gamma_2), \tag{13}$$

where $M_1^*(-\gamma_1) = (1-\tau^2)M_1(-\gamma_1) + \tau^2 I(s=0)$ is a convex combination of a valid covariance function and a nugget effect, rendering the expression in (13) a valid covariance function.

4 BAYESIAN MODEL AND INFERENCE

Consider that observations z_{ij} are obtained at locations s_i, $i = 1, \ldots, I$ and time points t_j, $j = 1, \ldots, J$. We confine ourselves to Gaussian joint distributions and the likelihood function is given by

$$l(\theta, \sigma^2, \mu; z) = (2\pi\sigma^2)^{-\frac{N}{2}}|\Sigma(\theta)|^{-1/2} \exp\left\{-\frac{1}{2}(\mathrm{Vec}(z) - \mu)^T \Sigma(\theta)^{-1}(\mathrm{Vec}(z) - \mu)\right\}, \tag{14}$$

with $N = IJ$ and $\Sigma(\theta)$ has elements

$$\Sigma(\theta)_{kk'} = C(s_k - s_{k'}, t_k - t_{k'}; \theta),\ k, k' = 1, \ldots, N,$$

where $C(s,t;\theta)$ is either of the covariance functions described in Subsection 3.1, possibly including a nugget effect as in (13). To complete the Bayesian model, we specific a prior on the parameter vector θ, which always contains $\sigma^2, a, b, \beta, \lambda_0, \lambda_1, \lambda_2$ as well as δ for Model 1 (while for Model 1a it contains δ but excludes λ_1). We consider independent priors in line with the more or less clear-cut different roles of the parameters. For the scale parameters in space and time we adopt gamma distributions, $a \sim \mathrm{Ga}(1, c_1/\mathrm{med}(d_s))$ and $b \sim \mathrm{Ga}(2,2)$. Here we have defined $\mathrm{med}(d_s)$ as the median spatial distance in the data, so that the prior on a takes into account the scaling of s. The prior distribution on θ was chosen in order to imply a reasonable prior distribution on c. We want a prior distribution for c that gives enough weight for values of c

close to 0 since this would mean a simpler (separable) model often used in the literature. This consideration lead us to use an Exponential prior with mean 1 for λ_0 in combination with $\lambda_2 \sim \mathrm{Ga}(3,1)$. In addition, for Model 1 we take $\lambda_1 \sim N(0,4), \delta \sim \mathrm{Ga}(1,1)$, for Model 2 $\nu = -\lambda_1 \sim \mathrm{Ga}(2,1)$ and for Model 3 we adopt $\lambda_1 \sim \mathrm{Ga}(2,1)$. In the case where λ_2 and/or λ_1 is fixed we suggest the use of an Exponential prior with mean $1/5$ for λ_0. The prior distribution for σ^{-2} is $\mathrm{Ga}(10^{-6}, 10^{-6})$, while we use a uniform distribution on $(0,2)$ for α and $\beta/2 \sim \mathrm{Beta}(3,2)$. For the models with nugget effect, the prior on τ^2 is $\mathrm{Ga}(2,6)$.

We need to be able to deal with a non-constant mean surface $\mu(s,t)$ in (14). Complex space-time trends are common in spatiotemporal data sets. For instance, temporal periodicity due to seasonal fluctuations may be combined with spatial trends due to geologic characteristics. The simplest case is when $\mu(s,t)$ is specified by a linear function of location, time and possible explanatory variables as follows

$$\mu(s,t) = \sum_{k=1}^{p} \psi_k f_k(s,t), \tag{15}$$

where $f_k(s,t),\ k = 1,\dots,p$ are functions of location, time or location and time, and $\psi = (\psi_1,\dots,\psi_p)'$ are unknown coefficients. As pointed out by Møller (2003, p. 54-56), linear or quadratic trend surfaces are useful descriptions of spatial mean but more complicated polynomials are seldom useful since they lead to unrealistic extrapolations beyond the region of observed locations. We suggest the prior $\psi|\sigma^2 \sim N_p(0, \sigma^2 V_0)$ with $V_0 = \sigma_0^2 I_p$, where σ_0^2 is large and I_p is the identity matrix.

We use stochastic simulation via MCMC to obtain an approximation of the posterior distribution of (θ, ψ). We obtain samples from the target distribution $p(\theta, \psi|z)$ by successive generations from the full conditional distributions. More specifically, we use a hybrid Gibbs sampler scheme with Metropolis-Hastings steps. The evaluation of the likelihood is the main computational requirement, and this is done through an efficient and accurate approximation as described in Appendix C. Examples with simulated data (for all models) confirm that the prior is not overly informative and that the numerical methods perform well.

Model comparison is conducted on the basis of Bayes factors. These are computed from the MCMC output using three different methods. We use the estimator p_4 of Newton and Raftery (1994) (with their d as small as 0.01), the optimal bridge sampling approach of Meng and Wong (1996), and the shifted gamma estimator proposed by Raftery et al. (2007) (with values of their λ close to one). Throughout, these three methods lead to very similar results.

5 APPLICATION: IRISH WIND DATA

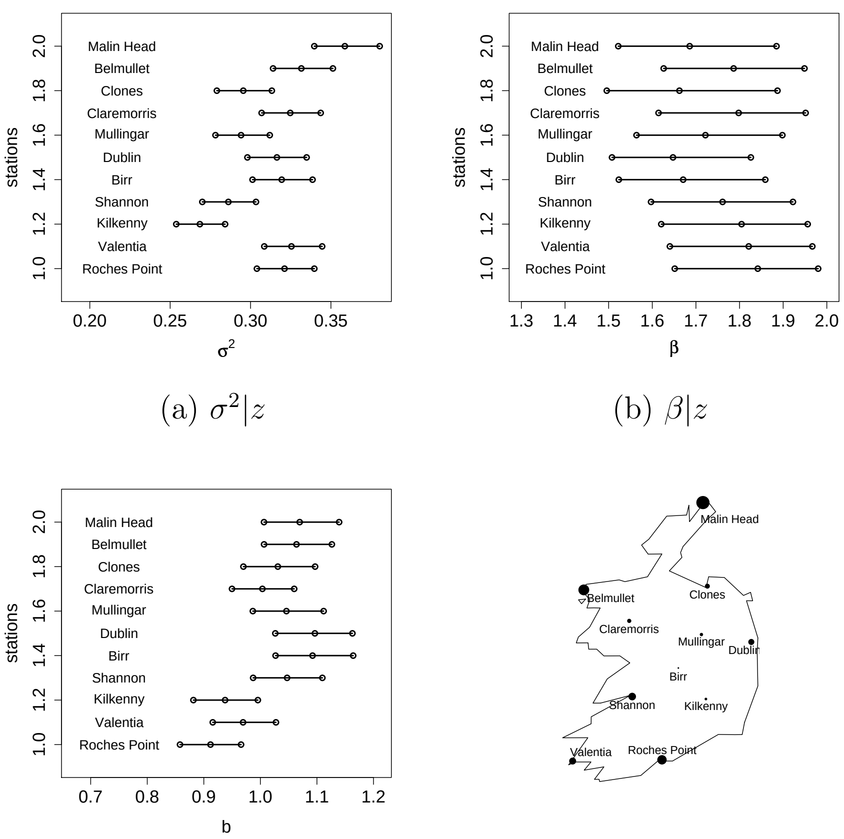
To illustrate the proposed models we analyze the Irish wind data described by Haslett and Raftery (1989) and also used in *e.g.* Li et al. (2007), Gneiting et al. (2007) and Stein (2005). The data consist of time series of daily average wind speed in m/s at 11 meteorological stations in Ireland during the period 1961-1978. We use UTM coordinates, so that the scale of the spatial coordinates is in kilometers and consider 10 years (1961-1970) of data. Following the literature we apply some transformations to the data. Firstly, we take the square root transformation in order to obtain data that are approximately normally distributed. Next, we estimate the seasonal effects by calculating the average of the square root of the daily means over years and stations for each day of the year and then regressing the result on the set of annual harmonics of the kind $(sin(\frac{2\pi}{365}rt), cos(\frac{2\pi}{365}rt)),\ t = 1, \ldots, 365$ and $r = 1, 2, \ldots, 364/2$. For this subset of the data we used $r = 1, 2, 3, 6$. We substract these estimated seasonal effects from the square root data and we work with the deseasonalized data.

The empirical correlation for the transformed wind speed decays fast in time, and for a lag of 4 days the empirical correlation is already very close to 0. Given the dimension of this data set we need to restrict our attention to few lags in time in order to make computation feasible. In this case, we use an approximation of the likelihood function which assumes that observations more than three days apart are uncorrelated. Without this simplification, it would be required to invert matrices of size $40,150 \times 40,150$ at each step of the MCMC algorithm in the spatiotemporal modeling. The details about the calculation of the likelihood in this example are presented in Appendix C.

In an initial analysis of the data set we fitted a purely temporal model station by station. We considered nonzero mean μ and we fitted a Cauchy covariance function given by

$$C_2(t; \sigma^2, \beta, b) = \sigma^2(1 + |t/b|^{\beta})^{-1},\ |t| \leq 10 \text{ days} \tag{16}$$

We consider Gaussian processes here, therefore for each station with location $s_i,\ i = 1, 2, \ldots, 11$ we have $z_i \sim N(\mu_i \iota, \Sigma_2)$, where z_i groups the 3650 observations for station i, ι is a vector of ones and $\Sigma_{2jj'} = C_2(t_j - t_{j'}; \sigma^2, \beta, b)$. Priors for the relevant parameters are as described in Section 4. Figure 3 shows the posterior medians as well as the 95% credible intervals for the parameters in (16) for each station. For all stations the correlation decays very rapidly in time. This preliminary analysis shows that the mean of the deseasonalized data still varies with site. This can be seen in Figure 3(d) which shows the posterior mean of the transformed wind speed $\mu_i,\ i = 1, \ldots, 11$ obtained from the purely temporal analysis for each station separately. The means are larger at the coastal



(a) $\sigma^2|z$ (b) $\beta|z$

(c) $b|z$ (d) $\mathrm{E}(\mu|z)$

Figure 3: (a-c) Posterior 95% credible intervals and median for the parameters in the temporal model for the 11 stations. (d) Posterior mean of μ for the 11 stations (the circle radius is proportional to the level).

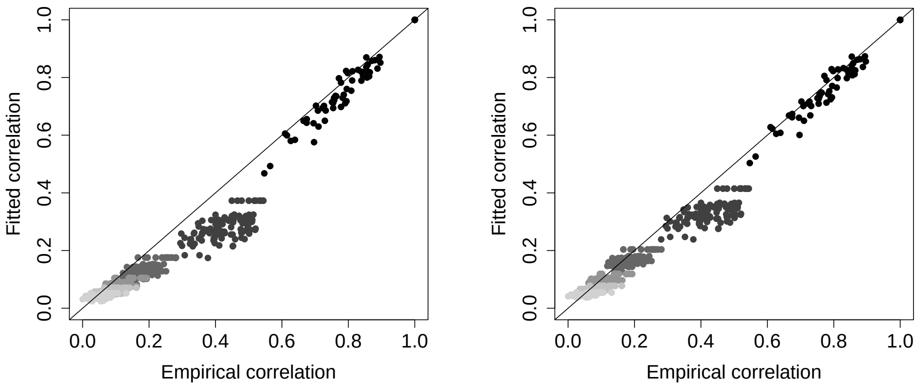
sites than inland and a sensible model for the mean should take this into consideration. The earlier papers that analyzed these data substract the mean by station or consider differences in time in order to avoid modeling this spatial trend. Since our goal is merely to compare models with different covariance structures, we remove the estimated (through posterior means) station-specific means. The resulting data are often called velocity measures. We also notice from this purely temporal analysis that the variances in the coastal sites are very different from the variances in the inland locations. We would like our model to capture this feature, therefore we consider different variances for each site. Stein (2005) comments that allowing for different variances at different sites can improve the fit considerably in this application.

In the spatiotemporal analysis we consider the proposed nonseparable model (8) with a constant $\mu(s,t)$ in (14). We set $\lambda_1 = -1$ and we estimate $\delta = a_1$,

α and a. As in the purely temporal analysis and as in earlier studies we set $\lambda_2 = 1$ for the Cauchy covariance in time and estimate b and β. We also want our model to capture discontinuities at the origin, so we include a purely spatial nugget effect as described in (13). Priors are as described in Section 4.

Parameter	50%	(2.5%, 97.5%)	50%	(2.5%, 97.5%)
τ^2	0.095	(0.087 , 0.103)	0.093	(0.084 , 0.101)
a	44.2	(30.6 , 62.1)	41.8	(29.9 , 49.2)
α	1.62	(1.49 , 1.78)	1.52	(1.40 , 1.69)
b	1.27	(1.02 , 1.75)	0.70	(0.68 , 0.73)
β	1.22	(1.11 , 1.33)	1.48	(1.42 , 1.54)
δ	0.07	(0.02 , 0.34)	0.79	(0.25 , 1.82)
λ_0	0.58	(0.31 , 1.03)	-	-

Table 1: Posterior median and 2.5% and 97.5% quantiles for the covariance parameters in the nonseparable and separable versions of Model 1.

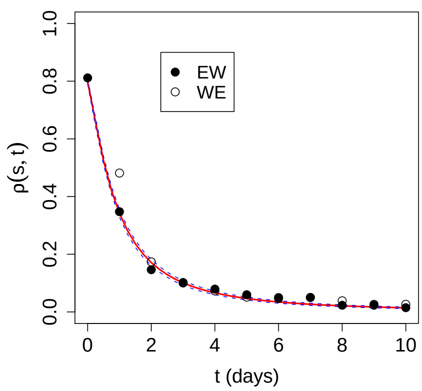


(a) Separable model (b) Nonseparable model

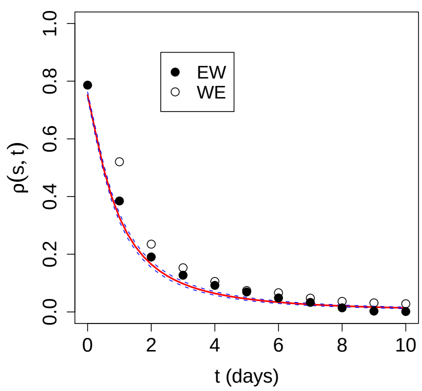
Figure 4: Empirical versus fitted (posterior median) correlations at temporal lags zero until five, with the higher lags corresponding to lighter shades, for the separable (a) and nonseparable (b) versions of Model 1.

We used a burn in of 10000 iterations and record every 18^{th} draw resulting in 5000 posterior draws from the posterior distribution. Table 1 shows the summaries from the posterior distribution for the nonseparable and separable versions of Model 1. The nugget effect is non-negligible and well estimated for both versions, while the posterior distributions of α and β suggests the process is not infinitely smooth in space and time. Note that the posterior mass for the separability parameter λ_0 is concentrated well away from zero, suggesting nonseparability.

The fitted (posterior median) correlation function versus the empirical correlation is presented in Figure 4 for the separable and nonseparable models (for all 55 pairs of locations). Clearly, the nonseparable model fits the empirical correlations better. The relative lack of fit at lag one in time is mainly due to the assumption of symmetry of the covariance function which is not adequate for this data set, as discussed in Li et al. (2007) and Gneiting et al. (2007). Figure 5 shows the difference between the empirical west-to-east (WE, *i.e.* with the westerly station leading in time) and east-to-west (EW, with the westerly station lagging in time) correlations and the fitted continuous correlation function using the nonseparable model for stations Valentia and Roche's Point (Figure 5(a)) and Belmullet and Clones (Figure 5(b)). The difference in the empirical correlation at lag one is quite large for both stations. A simple way to address this problem would be to consider $C(s - \epsilon t w, t)$ where ϵ is a parameter to be estimated and as the symmetries in this example are mainly functions of differences in longitude, Stein (2005) suggested $w = (0, 1)'$. But since an in-depth analysis of these data is not the main point of this paper, we proceed with a comparison between the various models proposed here.



(a) Valencia and Roche's Point



(b) Belmullet and Clones

Figure 5: Empirical EW and WE compared with fitted correlations (nonseparable Model 1). The drawn line is the posterior median and the dashed lines are the 95% credible intervals.

Table 2 shows the model comparison through Bayes factors.All the three methods mentioned in Section 4 for estimating the marginal likelihood give very strong evidence against the separable Model 1 in favor of its nonseparable version. Such support for the nonseparable model is also suggested by the posterior distribution of c in (7), which gives an measure of the degree of dependence in space and time (with $c = 0$ indicating complete separability). The 95% credible interval for c is $(0.218, 0.501)$ indicating strong nonseparability in this example. Figure 6 displays the posterior density of c, overlayed with its prior counterpart. Comparison of the nonseparable Model 1 with other nonseparable models indicates that Model 1 is also very strongly favored over the nonseparable exponential model (Model 2 with $\alpha = 2$ and $\nu = 0.5$) and decisively outperforms the nonseparable Cauchy model in space (Model 3).

	Newton-Raftery	Bridge-Sampling	Shifted-Gamma
Separable Model 1	49	46	50
Nonsep. Exponential	58	68	41
Nonsep. Cauchy	7	9	7

Table 2: The natural logarithm of the Bayes factor in favor of the nonseparable Model 1 versus the separable Model 1 and the nonseparable Models 2 and 3 using Newton-Raftery ($d = 0.01$), Bridge-sampling ($n = 1000$) and Shifted-Gamma ($\lambda = 0.98$) estimators for the marginal likelihood.

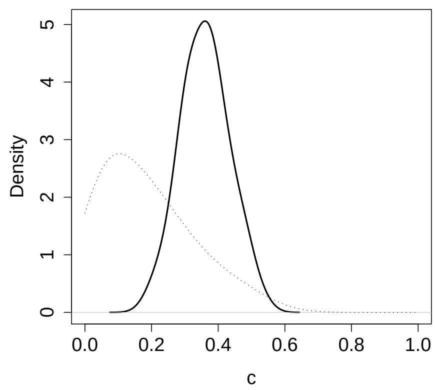


Figure 6: Posterior (solid line) and prior (dotted line) distributions for the c in (7).

6 CONCLUSION

In this article we have proposed a new covariance model that is nonseparable and includes the separable model as a particular case. The proposed model is obtained through a continuous mixture of separable covariances in space and time and it allows extensions (*e.g.* to include a nugget effect) in a very straightforward way. The resulting model has some useful theoretical properties such as different degrees of smoothness across space and time and long-range dependence in time. For practical modeling purposes, we suggest a number of different parameterisations, leading to a variety of special cases with a wide range of spatial behavior. Under separability, the purely spatial covariance structure can, for example, be a generalized Matérn, a Matérn, or a shifted exponential.

We conduct Bayesian inference with relatively vague priors, using an MCMC sampler. In addition, we implement an approximation to the likelihood which makes it feasible to perform inference for large data sets. We present an illustrative example for the Irish wind data of Haslett and Raftery (1989). The results show that despite the simplifying assumptions of stationarity and isotropy on which the model is based it is possible to obtain a very substantial improvement in the fit by using the nonseparable model proposed here.

Appendix A GIG Distribution

A random variable X has a generalized inverse gaussian (GIG) distribution if the density function of X is given by

$$f_{GIG}(x;\lambda,\delta,a) = \frac{(a/\delta)^{\lambda/2}}{2K_\lambda(2\sqrt{a\delta})} x^{\lambda-1}\exp\{-[ax+\delta x^{-1}]\},\ x>0. \tag{17}$$

Permitted parameter values are $a>0$, $\delta\geq 0$ if $\lambda>0$, while $a>0$, $\delta>0$ if $\lambda=0$ and $a\geq 0$, $\delta>0$ if $\lambda<0$. The standard reference for the GIG distribution is Jørgensen (1982). We use the notation $X\sim GIG(\lambda,\delta,a)$. An important aspect is that this class embraces many special cases such as the gamma distribution ($\lambda>0$ and $\delta=0$), the inverse gamma distribution ($\lambda<0$ and $a=0$), the inverse gaussian distribution ($\lambda=-1/2$) and the reciprocal inverse gaussian distribution ($\lambda=1/2$). For $a,\delta>0$, the mean of X is $\mathrm{E}(X)=\sqrt{\delta/a}\{K_{\lambda+1}(2\sqrt{a\delta})/K_\lambda(2\sqrt{a\delta})\}$ and the variance of X is given by

$$\mathrm{Var}(X) = \left(\frac{\delta}{a}\right)\left\{\frac{K_{\lambda+2}(2\sqrt{a\delta})}{K_\lambda(2\sqrt{a\delta})} - \frac{K^2_{\lambda+1}(2\sqrt{a\delta})}{K^2_\lambda(2\sqrt{a\delta})}\right\}. \tag{18}$$

Appendix B Proofs

Proof of Proposition 2.1

Consider the covariance model given in (3) and the specification $C_1(s;U)=\exp\{-\gamma_1 U\}$, $C_2(t;V)=\exp\{-\gamma_2 V\}$. If the random vector (U,V) has cumulative function $\mu(u,v)$ the joint moment generating function $M(r_1,r_2)=\int\exp\{r_1u+r_2v\}d\mu(u,v)$. Then,

$$C(s,t)=\int\exp\{-\gamma_1u-\gamma_2v\}d\mu(u,v)=M(-\gamma_1,-\gamma_2).$$

Proof of Proposition 3.1

From the conditions of Proposition 2.1 it follows that $C(s,t) = \sigma^2 M(-\gamma_1, -\gamma_2)$, where $M(.,.)$ is the joint moment generating function of (U,V). Define $U = X_0 + X_1$ and $V = X_0 + X_2$ with X_0, X_1 and X_2 independent non-negative random variables with finite moment generating function M_0, M_1 and M_2, respectively. Then,

$$\begin{aligned}
M(-\gamma_1, -\gamma_2) &= \mathrm{E}[\exp\{-\gamma_1 U - \gamma_2 V\}] \\
&= \mathrm{E}[\exp\{-(\gamma_1+\gamma_2)X_0 - \gamma_1 X_1 - \gamma_2 X_2\}] \\
&= M_0(-(\gamma_1+\gamma_2))M_1(-\gamma_1)M_2(-\gamma_2).
\end{aligned}$$

Proof of Theorem 3.1

Consider conditions of Proposition 3.1. Let $X_i \sim \mathrm{Ga}(\lambda_i, a_i)$, $i = 0, 2$, then

$$M_i(r) = \mathrm{E}[\exp\{rX_i\}] = \int_0^\infty \frac{a_i^{\lambda_i}}{\Gamma(\lambda_i)} x^{\lambda_i - 1}\exp\{-(a_i - r)x\}dx = \left(\frac{a_i}{a_i - r}\right)^{\lambda_i}, \ r < a_i.$$

Let $X_1 \sim \mathrm{GIG}(\lambda_1, \delta, a_1)$, if $a_1 \neq 0$ then

$$\begin{aligned}
M_1(r) &= \mathrm{E}[\exp\{rX_1\}] \\
&= \int_0^\infty \frac{(\delta/a_1)^{-\lambda_1/2}}{2K_{\lambda_1}(2\sqrt{\delta a_1})} x^{\lambda_1 - 1}\exp\{-[(a_1 - r)x + \delta x^{-1}]\}dx \\
&= \left(\frac{a_1}{a_1 - r}\right)^{\lambda_1/2} \frac{K_{\lambda_1}(2\sqrt{(a_1 - r)\delta})}{K_{\lambda_1}(2\sqrt{\delta a_1})}, \ r < a_1,
\end{aligned}$$

if $a_1 = 0$ we use the asymptotic formula $K_{\lambda_1}(2\sqrt{\delta a_1}) = 2^{\lambda_1 - 1}\Gamma(\lambda_1)(2\sqrt{\delta a_1})^{-\lambda_1}$ implying

$$M_1(r) = \left(\frac{a_1}{a_1 - r}\right)^{\frac{\lambda_1}{2}} \frac{K_{\lambda_1}(2\sqrt{-r\delta})}{2^{\lambda_1-1}\Gamma(\lambda_1)(2\sqrt{\delta a_1})^{-\lambda_1}} = \frac{(2\sqrt{-r\delta})^{\lambda_1}}{\Gamma(\lambda_1)2^{\lambda_1 - 1}} K_{\lambda_1}(2\sqrt{-r\delta}), \ r < 0, \tag{19}$$

Theorem 3.1 follows.

Proof of Proposition 3.2

(a) The covariance function for the process $\{Z(s_0, t) : t \in T\}$ at a fixed location $s_0 \in D$ is given by

$$C(0,t) = \sigma^2 M_0(-\gamma_2(t))M_2(-\gamma_2(t)) \tag{20}$$

where $M_i(r) = \left\{1 - \frac{r}{a_i}\right\}^{-\lambda_i}$, $r < a_i$, $i = 0, 2$ and $a_0 = 1$. A (weakly) stationary process with covariance function $K(t)$ is m times mean square differentiable if and only if $K^{(2m)}(0)$ exists and is finite (see Stein (1999) pp 20-22). By Faá di Bruno's formula, termwise differentiation of (20) results in

$$\begin{aligned}
C^{(2m)}(0,t) &= \sigma^2 \sum_A \frac{m!}{k_1!k_2!...k_{2m}!} y^{(k)}(-\gamma_2(t)) \prod_{k_j \neq 0} \left\{\frac{-\gamma_2^{(j)}(t)}{j!}\right\}^{k_j} \qquad (21) \\
&= \sigma^2\{y^{(1)}(-\gamma_2(t))[\gamma_2^{(2m)}(t)] + \ldots + y^{(2m)}(-\gamma_2(t))[\gamma_2^{(1)}(t)]^{2m}\}
\end{aligned}$$

where $A = \{k_1, k_2, ..., k_{2m} : k_1+2k_2+...+2mk_{2m} = 2m\}$, $k = k_1+k_2+...+k_{2m}$, $k_i \geq 0,\ i = 1, 2, ..., 2m$ and

$$y^{(k)}(r) = \sum_{i=0}^{k} \binom{k}{i} M_0^{(k-i)}(r) M_2^{(i)}(r).$$

The terms $y^{(1)}(-\gamma_2(t)), \ldots, y^{(2m)}(-\gamma_2(t))$ exist and are finite as $t \to 0$ since

$$y^{(k)}(0) = \sum_{i=0}^{k} \binom{k}{i} \mathrm{E}(X_0^{k-i})\mathrm{E}(X_2^i) = \sum_{i=0}^{k} \binom{k}{i} \frac{\Gamma(\lambda_0+k-i)}{\Gamma(\lambda_0)} \frac{\Gamma(\lambda_2+i)}{\Gamma(\lambda_2)} \left(\frac{1}{a_2}\right)^i,$$

where $k = 1, 2, \ldots, 2m$. In the expression (21), the highest order derivative of $\gamma_2(t)$ is $2m$ obtained when $k_{2m} = 1$ and $k_1 = \ldots = k_{2m-1} = 0$. Thus the behavior of $C^{(2m)}(0, t)$ as $t \to 0$ depends only on the local behavior of $\gamma_2^{(2m)}(t)$ as $t \to 0$.

(b) The covariance function for the process $\{Z(s, t_0) : s \in D\}$ at a fixed time $t_0 \in T$ is given by

$$C(s, 0) = \sigma^2 M_0(-\gamma_1(s)) M_1(-\gamma_1(s)), \tag{22}$$

$M_0(r) = \{1-r\}^{-\lambda_0},\ r < 1$ and $M_1(r) = \left\{1 - \frac{r}{a_1}\right\}^{-\frac{\lambda_1}{2}} \frac{K_{\lambda_1}(2\sqrt{(a_1-r)\delta})}{K_{\lambda_1}(2\sqrt{a_1\delta})},\ r < a_1$. By Faá di Brunos's formula, termwise differentiation of (22) results in

$$\begin{aligned} C^{(2m)}(s, 0) &= \sigma^2 \sum_A \frac{m!}{k_1!k_2!...k_{2m}!} y^{(k)}(-\gamma_1(s)) \prod_{k_j \neq 0} \left\{ \frac{-\gamma_1^{(j)}(s)}{j!} \right\}^{k_j} \\ &= \sigma^2 \{ y^{(1)}(-\gamma_1(s))[\gamma_1^{(2m)}(s)] + \ldots + y^{(2m)}(-\gamma_1(s))[\gamma_1^{(1)}(s)]^{2m} \} \end{aligned} \tag{23}$$

where $A = \{k_1, k_2, ..., k_{2m} : k_1 + 2k_2 + ... + 2mk_{2m} = 2m\}$, $k = k_1 + k_2 + ... + k_{2m}$, $k_i \geq 0$, $i = 1, 2, \ldots, 2m$ and $y^{(k)}(x) = \sum_{i=0}^{k} \binom{k}{i} M_0^{(k-i)}(x) M_1^{(i)}(x)$.

(i) Consider $a_1 \neq 0$. The terms $y^{(1)}(-\gamma_2(t)), \ldots, y^{(2m)}(-\gamma_2(t))$ exist and is finite for all integer m as $s \to 0$ since

$$y^{(k)}(0) = \sum_{i=0}^{k} \binom{k}{i} \mathrm{E}(X_0^{k-i})\mathrm{E}(X_1^i) = \begin{cases} \sum_{i=0}^{k} \binom{k}{i} \frac{\Gamma(\lambda_0+k-i)}{\Gamma(\lambda_0)} \frac{\Gamma(\lambda_1+i)}{\Gamma(\lambda_1)} \left(\frac{1}{a_1}\right)^i & \text{if } \delta = 0 \\ \sum_{i=0}^{k} \binom{k}{i} \frac{\Gamma(\lambda_0+k-i)}{\Gamma(\lambda_0)} \frac{K_{\lambda_1+i}(2\sqrt{a_1\delta})}{K_{\lambda_1}(2\sqrt{a_1\delta})} \left(\frac{\delta}{a_1}\right)^{i/2} & \text{if } \delta \neq 0 \end{cases}$$

$k = 1, 2, \ldots, 2m$. In the expression (23), the highest order derivative of $\gamma_1(s)$ is $2m$ obtained when $k_{2m} = 1$ and $k_1 = \ldots = k_{2m-1} = 0$. Thus the behavior of $C^{(2m)}(s, 0)$ as $s \to 0$ depends only on the local behavior of $\gamma_1^{(2m)}(s)$ as $s \to 0$.

(ii) Consider $a_1 \neq 0$ (which implies $\lambda_1 < 0$). We need to study the behavior of

$$y^{(1)}(-\gamma_1(s))[-\gamma_1^{(2m)}(s)] \tag{24}$$

obtained when $k_{2m} = 1$ and $k_1 = \ldots = k_{2m-1} = 0$ and

$$y^{(2m)}(-\gamma_1(s))[-\gamma_1^{(1)}(s)]^{2m} \tag{25}$$

obtained when $k_1 = 2m$ and $k_2 = \ldots = k_{2m} = 0$ as the other terms in (23) require lower order derivatives.

Consider the Taylor expansion of $\gamma_1(s)$ about 0 given by $\gamma_1(s) = \sum_{j=0}^{\infty} c_j ||s||^j$. Thus the behavior of $\gamma_1(s)$ as $s \to 0$ is determined by $||s||^l$ where l is the smallest power such that $c_l \neq 0$.

The expression (25) is given by

$$y^{(2m)}(-\gamma_1(s))[-\gamma_1^{(1)}(s)]^{2m} = \sum_{i=1}^{2m} \binom{2m}{i} M_0^{(2m-i)}(-\gamma_1(s)) M_1^{(i)}(-\gamma_1(s))[-\gamma_1^{(1)}(s)]^{2m}$$

and the highest order derivative of $M_1(-\gamma_1(s))$ is $2m$ obtained when $i = 2m$. Thus (25) will exist and be finite if $||s||^{l(-\lambda_1-2m)} \times ||s||^{(l-1)2m} = ||s||^{(-l\lambda_1-2m)}$ exists and is finite, implying that m need to satisfy $2m < -l\lambda_1$.

The expression (24) is given by

$$M_0^{(1)}(-\gamma_1(s))M_1(-\gamma_1(s))[-\gamma_1^{(2m)}(s)] + M_0(-\gamma_1(s))M_1^{(1)}(-\gamma_1(s))[-\gamma_1^{(2m)}(s)]$$

and as $s \to 0$ we obtain $-\mathrm{E}[X_0]\gamma_1^{(2m)}(s) + ||s||^{l(-\lambda_1-1)}||s||^{l^*-2m}$, where $l^* \geq l$ implying that is is sufficient for m to satisfy $2m < -l\lambda_1$ and $\gamma_1^{(2m)}(0)$ needs to exist and be finite. Thus, the purely spatial process is m times mean square differentiable if and only if $\gamma_1^{(2m)}(s)$ as $s \to 0$ exists and is finite and $2m < -l\lambda_1$.

Appendix C Computational issues

Consider the matrix Z with elements $Z_{ij} = Z(s_i, t_j)$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$. We split the data into $Z_1, Z_2, \ldots, Z_{J/L}$ each of size $I \times L$ and L is the lag in time (we assume, for simplicity, that J/L is integer, but the code can deal with the general case by simply adapting the dimension of the last block). We will approximate the likelihood by assuming that the temporal correlation is positive only between observations that are not more than L time periods apart and that it is zero otherwise. This is justified by the very small empirical correlation at time lags larger than L. For the Irish wind data we take $L = 3$ and taking larger values of L did not affect the results appreciably. Consider $y_k = \mathrm{Vec}(Z_k)$, where the first I elements correspond to observations for all locations at time $(k-1)L+1$, etc., then the covariance matrix of $(y_1, \ldots, y_{J/L})$ has a block Toeplitz structure given by

$$\begin{pmatrix} T_{11} & T_{12} & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ T_{21} & T_{11} & T_{12} & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & T_{21} & T_{11} & T_{12} & \ldots & 0 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & 0 & T_{21} & T_{11} & T_{12} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & T_{21} & T_{11} \end{pmatrix}$$

where

$$T_{11} = \begin{pmatrix} \Sigma^{(0)} & \Sigma^{(1)} & \dots & \Sigma^{(L-2)} & \Sigma^{(L-1)} \\ \Sigma^{(1)} & \Sigma^{(0)} & \dots & \Sigma^{(L-3)} & \Sigma^{(L-2)} \\ & & & & \\ \Sigma^{(L-2)} & \Sigma^{(L-3)} & \dots & \Sigma^{(0)} & \Sigma^{(1)} \\ \Sigma^{(L-1)} & \Sigma^{(L-2)} & \dots & \Sigma^{(1)} & \Sigma^{(0)} \end{pmatrix}$$

and

$$T_{12} = \begin{pmatrix} \Sigma^{(L)} & 0 & \dots & 0 & 0 \\ \Sigma^{(L-1)} & \Sigma^{(L)} & \dots & 0 & 0 \\ & & & & \\ \Sigma^{(2)} & \Sigma^{(3)} & \dots & \Sigma^{(L)} & 0 \\ \Sigma^{(1)} & \Sigma^{(2)} & \dots & \Sigma^{(L-1)} & \Sigma^{(L)} \end{pmatrix}$$

where $\Sigma_{ij}^{(t)} = C(s_i - s_j, t)$, $i, j = 1, \dots, I$. Notice that in our example the time is equally spaced with intervals of one day. Then, we calculate the likelihood using the factorization

$$p(y|\theta) = p(y_1|\theta)p(y_2|y_1, \theta) \dots p(y_{\frac{J}{L}}|y_1, \dots, y_{\frac{J}{L}-1}, \theta)$$

where $y_1|\theta \sim N(\mu_1 = 0, V_1 = T_{11})$ and $y_k|y_1, \dots, y_{k-1}, \theta \sim N(\mu_k, V_k)$. $\mu_k = T_{21}V_{k-1}^{-1}(y_{k-1} - \mu_{k-1})$ and $V_k = T_{11} - T_{21}V_{k-1}^{-1}T_{12}$ for $k = 2, \dots, \frac{J}{L}$.

In our example we have relatively few stations and many replications in time. Therefore the critical issue is how to treat the time dimension in the calculation of the likelihood. The strategy chosen is particularly advantageous in this situation of stationarity and isotropy because we just need to invert small matrices with dimension $IL \times IL$.

References

Carrol, R. J., Chen, R., George, E. I., Li, T. H., Newton, H. J., Schmiediche, H., and Wang, N. (1997). "Ozone Exposure and Population Density in Harris County, Texas." *Journal of the American Statistical Association*, 92, 438, 392–404.

Chilès, J.-P. and Delfiner, P. (1999). *Modeling Spatial Uncertainty*. New York: Wiley.

Cressie, N. (1997). "Comment on Carrol et al. (1997)." *Journal of the American Statistical Association*, 92, 411–413.

Cressie, N. and Huang, H.-C. (1999). "Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions." *Journal of the American Statistical Association*, 94, 448, 1330–1340.

De Cesare, L., Myers, D. E., and Posa, D. (2001). "Estimating and Modeling Space-Time Correlation Structures." *Statistics and Probability Letters*, 51, 9–14.

Gneiting, T. (2002). "Nonseparable, Stationary Covariance Functions for Space-Time Data." *Journal of the American Statistical Association*, 97, 458, 590–600.

Gneiting, T., Genton, M. G., and Guttorp, P. (2007). *Statistical Methods for Spatio-Temporal Systems*, vol. 107 of *Monographs on Statistics and Applied Probability*, chap. Geostatistical space-time models, stationarity, separability and full symmetry, 151–175. Chapman and Hall.

Gneiting, T. and Schlather, M. (2004). "Stochastic Models That Separate Fractal Dimension and the Hurst Effect." *SIAM review*, 46, 269–282.

Guttorp, P., Meiring, W., and Sampson, P. D. (1997). "Comment on Carrol et al. (1997)." *Journal of the American Statistical Association*, 92, 405–408.

Haslett, J. and Raftery, A. E. (1989). "Space-Time Modeling With Long-Memory Dependence: Assessing Ireland's Wind-Power Resource." *Applied Statistics*, 38, 1–50.

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*, vol. 9 of Lecture Notes in Statistics. Heidelberg: Springer.

Li, B., Genton, M. G., and Sherman, M. (2007). "A Nonparametric Assessment of Properties of Space-Time Covariance Functions." *Journal of the American Statistical Association*, 102, 736–744.

Ma, C. (2002). "Spatio-Temporal Covariance Functions Generated by Mixtures." *Mathematical geology*, 34, 8, 965–975.

— (2003). "Spatio-Temporal Stationary Covariance Models." *Journal of Multivariate analysis*, 86, 97–107.

Meng, X. L. and Wong, W. H. (1996). "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration." *Statistica Sinica*, 6, 831–860.

Møller, J. (2003). *Spatial Statistics and Computational Methods*. Springer.

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap." *Journal of the Royal Statistical Society. Series B*, 56, 3–48.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). "Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity." In *Bayesian Statistics 8*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 371–416. Oxford: Oxford University Press.

Shkarofsky, I. P. (1968). "Generalized Turbulence Space-Correlation and Wave-Number Spectrum-Function Pairs." *Canadian Journal of Physics*, 46, 2133–2153.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer New York.

— (2005). "Space-Time Covariance Functions." *Journal of the American Statistical Association*, 100, 469, 310–321.